# **EXPERIMENT REPORT 3**

|  |  |
| --- | --- |
| **Student Name** | Ngoc Quang Vinh Pham |
| **Project Name** | Predicting Customer Car Repurchase Likelihood |
| **Date** | April 14, 2024 |
| **Deliverables** | 36106-AT2-25100660-experiment-3.ipynb  Model Comparison & Evaluation Metrics Examination |

|  |  |
| --- | --- |
| 1. **EXPERIMENT BACKGROUND** | |
| Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach. | |
| **1.a. Business Objective** | This study seeks to predict which consumers are most likely to make another automobile purchase, which will improve the accuracy of focused marketing activities. By concentrating efforts on clients who have a greater tendency to purchase, accurate forecasts may dramatically save marketing expenses and boost sales efficiency. On the other side, inaccurate outcomes might result in lost opportunities or resource waste. |
| **1.b. Hypothesis** | The experiment's hypothesis is that the effectiveness of various machine learning models and assessment metrics in forecasting consumer behavior will differ. The False Positive rate must be minimized more so than the False Negative. In other words, wrongly targeting a customer who is less likely to repurchase is costlier than missing out on one who does.We compare various models and metrics in an effort to determine which combination works most effectively in forecasting car repurchases. |
| **1.c. Experiment Objective** | Comparing the effectiveness of four machine learning models—XGBoost, Random Forest, KNN, and SVM—using four distinct assessment metrics—F1, F2 (which emphasizes recall), F3 (which emphasizes recall even more), and ROC-AUC—is the goal. For our particular business adjusting, we want to figure out which model and metric combination yields the most accurate predictions. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT DETAILS** | |
| Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them. | |
| **2.a. Data Preparation** | Gender Binary Encoding: We encoded the gender variable as binary (Male = 1, Female = 0), facilitating its use in our models. To preserve the integrity of the dataset and prevent bias introduction, missing values in gender were imputed using the current distribution (58% Male).  Dropping ‘age\_band’: This feature was removed due to the high prevalence of missing values in ‘age\_band’ (85%). The rationale for this was that imputing such a vast amount of missing data could lead to some significant inaccuracies and unreliable predictions.  One-hot encoding: Since there is only 4 unique values in ‘car-segment’, I decided to encode ‘car-segment’ into multiple binary variables. This method is required to handle categorical data, ensuring that the model correctly interprets these categories as distinct without any inherent order.  ‘Car\_model’ Frequency Encoding: To handle the large correlation of this feature, car\_model received frequency encoding. Compared to one-hot encoding, this method reduces dimensionality while preserving information regarding model popularity, which may be indicative of repurchase behavior. |
| **2.b. Feature Engineering** | In this stage, we did not add new characteristics to our dataset. This choice was made to first evaluate the baseline predictive potential of existing characteristics without adding complexity. Features that have more than 60% null values will be removed from the dataset due to the possibility of inaccurate and incorrect imputation of these characteristics.  In the future, demographic information and history purchasing behavior interaction aspects may provide insights. Such characteristics may reveal patterns that individual features cannot. To evaluate non-linear correlations between characteristics and the target variable, polynomial features for continuous variables might be examined.  Segmenting clients by their marketing campaign interactions may reveal tendencies that affect repurchase probability.  By applying train\_test\_split from library sklearn, I have split the data into 3 different datasets:   * Training dataset: 91935 records (70%) * Validation dataset: 19701 records (15%) * Testing dataset: 19701 records (15%)   The scaling method selected for this project is the Standard Scaler, based on the results from the previous experiment.  The Standard SMOTE technique is implemented in our training dataset in order to enhance the complexity of the models by synthetically over-sampling minority data and then apply the trained model to the unseen validation and testing dataset with a highly imbalanced distribution. |
| **2.c. Modelling** | In this experiment, I choose 4 different machine learning models that have the highest performances from the last experiment. They are:   * Extreme Gradient Boosting Classifier (XGB) * Random Forest Classifier (RF) * Support Vector Machine (SVM) * K-Nearest Neighbors (KNN)   These models were selected because of their diverse methodologies of addressing binary classification problems, which will help us in evaluating the effectiveness of each model under various evaluation metrics. As I mention in the previous section, Standard SMOTE was also applied to address the class imbalance in the training set. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT RESULTS** | |
| Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified. | |
| **3.a. Technical Performance** | The F1, F2, F3, and ROC-AUC scores were used to assess each model's ability to manage the trade-off between precision and recall:   * F1-score: evaluate the balance between precision and recall. * F2-score: put more weight on recall, useful in scenarios where missing a positive instance has greater consequences. * F3-score: give much more weight to the recall making it appropriate for highly sensitive cases. * ROC-AUC-score: measure the model’s capacity to distinguish between classes at different threshold values.   The XGB model demonstrated superior technical performance across all metrics, indicating a robust capability to balance precision and recall and adjust to a recall-oriented focus (F2, F3). Random Forest followed closely, excelling particularly in ROC-AUC, suggesting strong discrimination abilities. KNN underperformed across the board, suggesting it may be less suited for the data's complexities. SVM showed adequate discrimination but lower precision-recall balance. |
| **3.b. Business Impact** | In marketing tactics where failure to identify a possible repurchaser might result in considerable opportunity costs, models with high F2 and F3 scores would be very beneficial. Precise predictions based on the experiment have the ability to improve sales and customer retention by optimizing marketing initiatives and focusing resources on clients who are most likely to make further purchases. However, inaccurate forecasts might result in missed sales opportunities from untargeted probable consumers and wasted marketing resources on low-probability clients. The models' ability to discriminate and generalize were shown using the ROC-AUC measure. |
| **3.c. Encountered Issues** | Bias Towards the Majority Class: Initial model assessments revealed a bias in favor of the majority class, which might be problematic in datasets that are unbalanced and where the minority class is more interesting.  Overfitting: Recall and accuracy have to be balanced. Recall overemphasis may result in overfitting, a situation in which the model performs well on training data but badly on untrained data.  Computational Efficiency: KNN and SVM required significant computation time, especially with large datasets and when used with SMOTE. |

|  |  |
| --- | --- |
| 1. **FUTURE EXPERIMENT** | |
| Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective. | |
| **4.a. Key Learning** | This experiment illustrated how crucial it is to use appropriate evaluation metrics that are relevant to certain business requirements. However, as the anticipation of the ‘Target’ value ‘1’ as the number of customers that have purchased more than 1 vehicle is more importance when these customers have high probability of new purchase. For these reasons, the False Positive is more significant than the False Negative which means the higher Recall value, the better the model. Nevertheless, the harmonic between Precision and Recall should remain the same to avoid overfitting problem. Hence, the F3-Score might be the best evaluation metrics for the assessment of the model performance.  Overall, XGBoost and Random Forest performed better, indicating that further fine-tuning and testing with these models may be beneficial. |
| **4.b. Suggestions / Recommendations** | Further Model Tuning: To improve model performance, investigate ensemble methods and more sophisticated hyperparameter tuning.  Hybrid Models: To increase precision and adaptability, think about combining model predictions.  Deployment Strategy: Create a strategy for the best-performing model's deployment that includes continuing monitoring and upgrading systems.  With the use of this study, marketing analytics prediction models may be improved, resulting in more effective and efficient targeted initiatives. |